Xgboost for station-level-prediction

Jose Peeterson

**Table of Contents**

[a) Clean-up the ‘xgb\_stn\_dem\_pred’ branch in DeepAR\_demand\_prediction repository. 3](#__RefHeading___Toc220_3406738982)

[b) Requirements and Architecture 3](#__RefHeading___Toc216_3406738982)

[Functionality requirements: 3](#__RefHeading___Toc332_3406738982)

[1) List of existing features: 3](#__RefHeading___Toc323_3786217518)

[2) Proposed new features: 3](#__RefHeading___Toc325_3786217518)

[3) Type of encoding of categorical features: 3](#__RefHeading___Toc327_3786217518)

[4) Automatic feature generation packages: 4](#__RefHeading___Toc329_3786217518)

[Data ingestion: 4](#__RefHeading___Toc334_3406738982)

[Loss function: 4](#__RefHeading___Toc336_3406738982)

[Metrics: 4](#__RefHeading___Toc338_3406738982)

[1) Exact prediction: 4](#__RefHeading___Toc348_3406738982)

[2) Leading prediction: 4](#__RefHeading___Toc350_3406738982)

[3) Error visualization: 5](#__RefHeading___Toc352_3406738982)

[4) Feature importance: 5](#__RefHeading___Toc372_3406738982)

[c) Feature Analysis study 5](#__RefHeading___Toc218_3406738982)

[Trial #1: 5](#__RefHeading___Toc354_3406738982)

[1) Dataset used: 5](#__RefHeading___Toc374_3406738982)

[2) Features Importance: 5](#__RefHeading___Toc356_3406738982)

[3) Loss: 7](#__RefHeading___Toc376_3406738982)

[4) Best hyperparameters: 7](#__RefHeading___Toc378_3406738982)

[5) Results: 7](#__RefHeading___Toc358_3406738982)

[Precision,Recall & F1 score 9](#__RefHeading___Toc317_3786217518)

[7) Analysis: 9](#__RefHeading___Toc360_3406738982)

[Trial #2: More features and more region level dataset 10](#__RefHeading___Toc362_3406738982)

[1) Dataset used: 10](#__RefHeading___Toc380_3406738982)

[2) Features Importance: 10](#__RefHeading___Toc364_3406738982)

[3) Results: 10](#__RefHeading___Toc366_3406738982)

[Precision,Recall & F1 score 10](#__RefHeading___Toc319_3786217518)

[4) Analysis: 10](#__RefHeading___Toc368_3406738982)

# Clean-up the ‘xgb\_stn\_dem\_pred’ branch in DeepAR\_demand\_prediction repository.

**15/4/23**

**There are some files deleted in other branches when xgb\_stn\_dem\_pred is files are deleted. between**

# Requirements and Architecture

## Functionality requirements:

### **List of existing features:**

* 1. DeepAR\_agg\_outflow
  2. rem\_blk\_outf
  3. net\_inflow\_stn
  4. en\_route\_inf
  5. net\_inflow\_clstr
  6. p\_1wk\_o\_
  7. p\_2wk\_o\_
  8. p\_3wk\_o\_
  9. p\_1ts\_o\_
  10. p\_2ts\_o\_
  11. p\_3ts\_o\_
  12. block\_id
  13. ts\_of\_day\_
  14. hr\_of\_day
  15. day\_of\_wk
  16. day\_of\_mn
  17. wk\_of\_mon

### Proposed new features:

1. tsfresh features
2. differencing time series
3. latitude and longitude
4. future (next time slot) inflow

### Type of encoding of categorical features:

1. <https://www.kaggle.com/code/subinium/11-categorical-encoders-and-benchmark>
2. one-hot encoding
3. target encoding

### Automatic feature generation packages:

1. Entire list
   * 1. tsfresh
     2. tsfeatures
     3. feature-engine
     4. cesium

## Data ingestion:

1. **size of data**
   1. Existing:
      1. All stations within a cluster

* 1. Proposed:
     1. All stations within a region

## Loss function:

1. **Regression:**
   1. Tweedie log likelihood
2. **Classification**
   1. cross-entropy loss (from softmax output for multi-class)

## Metrics:

### Exact prediction:

1. **Regression:**
   1. Tweedie log likelihood
   2. Precision and Recall of individual classes after rounding to Integer
   3. F1 score for individual classes
2. **Classification**
   1. cross-entropy loss (from softmax output)
   2. Precision and Recall of individual classes
   3. F1 score for individual classes

### Leading prediction:

METRIC to quantify leading predictions (predicted in a 10/20 minute time slot before current time slot)

1. First, the input features should have some future predictor variables like weather in next 10 mins And/Or inflow in next 10 mins. In order to justify the possibility of a non-random leading prediction.
2. How to compute this metric?
3. Remove all correct predictions to leave misclassified predictions
4. Check the prediction from 10 min leading time slot and compare with actual of current time slot.
5. If they are equal count 0. if they are unequal count 1.
6. Fraction of total unequal to total misclassified is the metric.

### Error visualization:

1. **Confusion matrix**
2. **Density scatter plot**
3. **Normal plot**

### Feature importance:

1. **plot bar chart of feature importance**

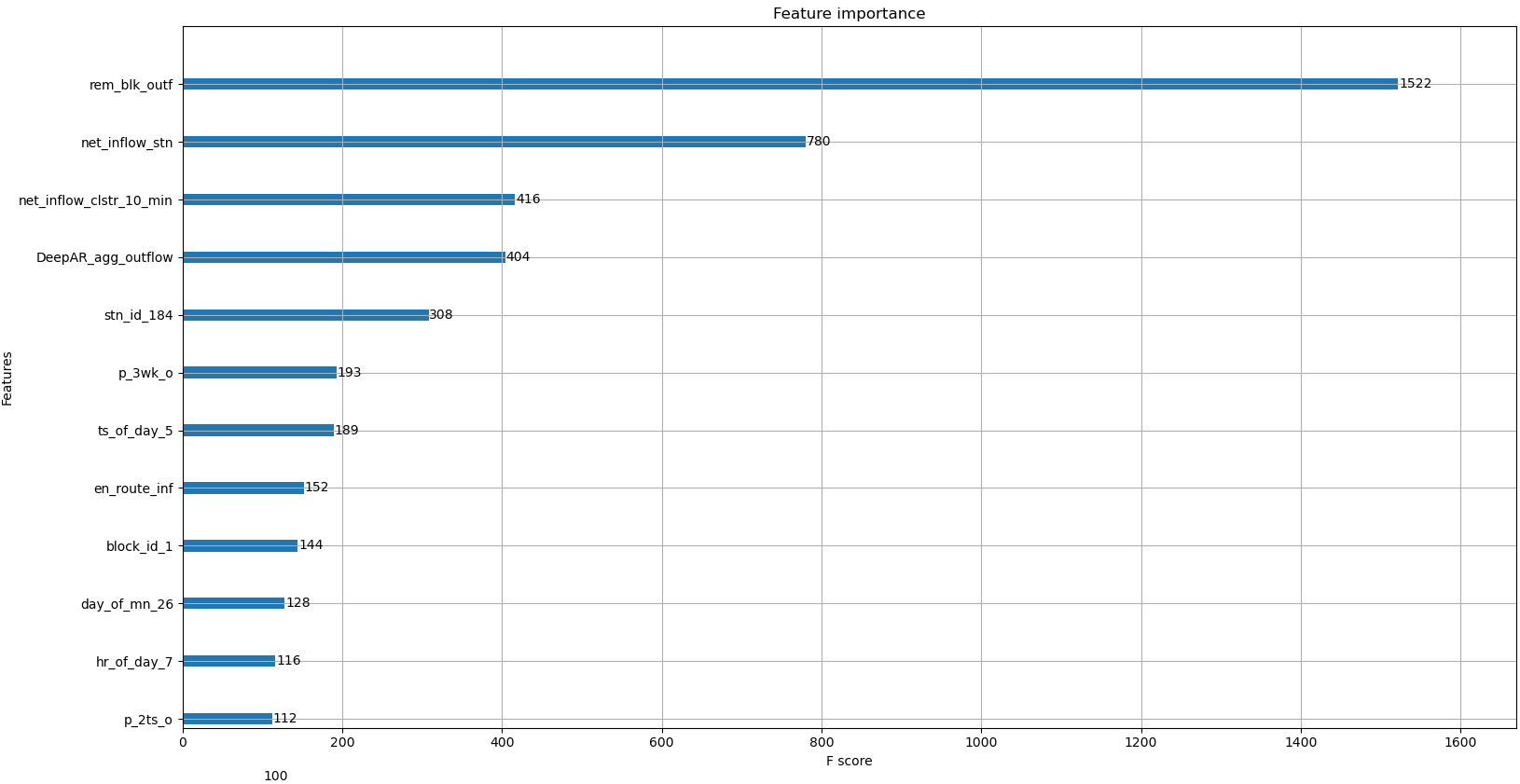
# Feature Analysis study

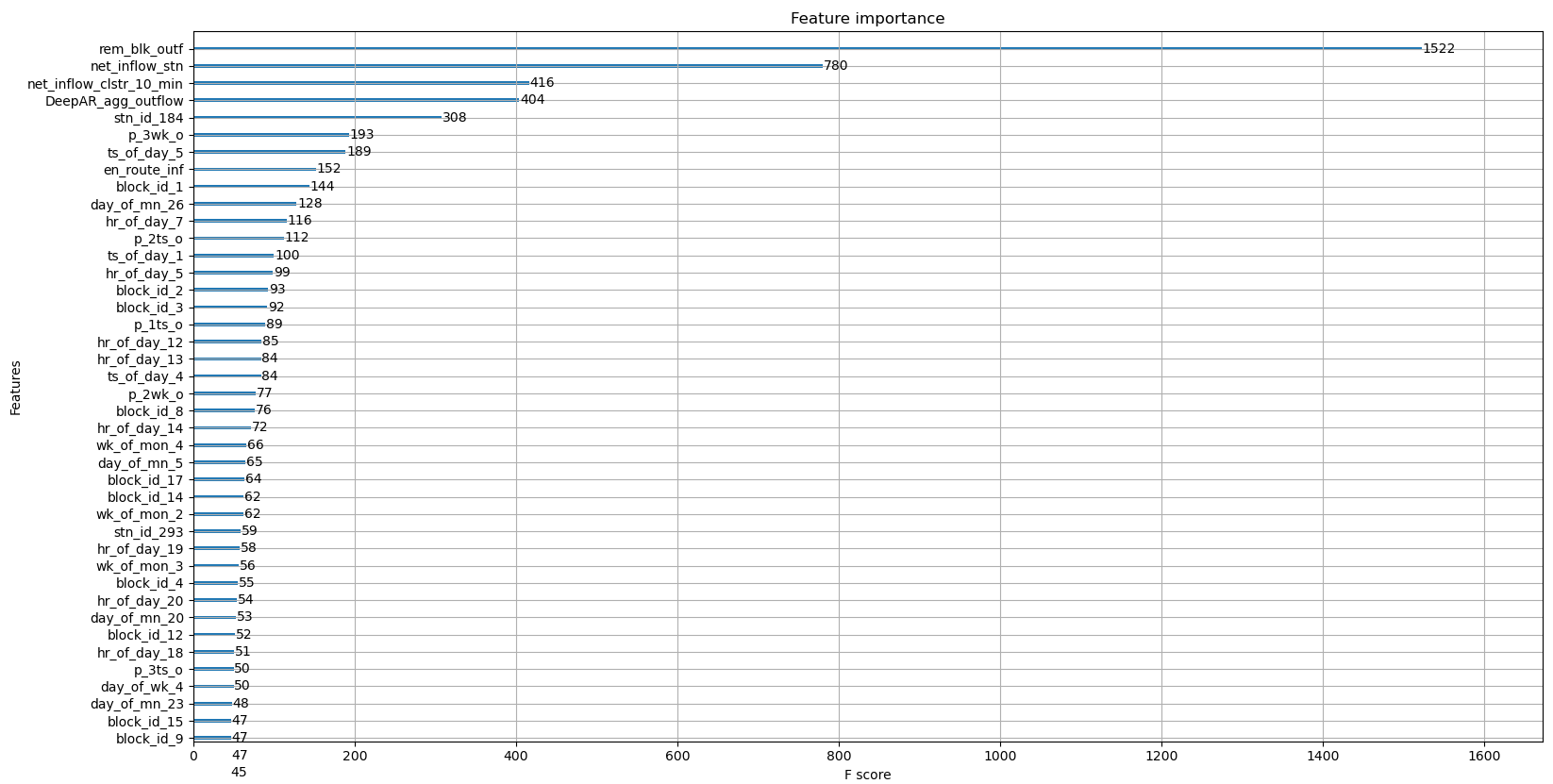
## Trial #1: Tweedie regression

### Dataset used:

All stations of one cluster.

### Features Importance:





### Loss:

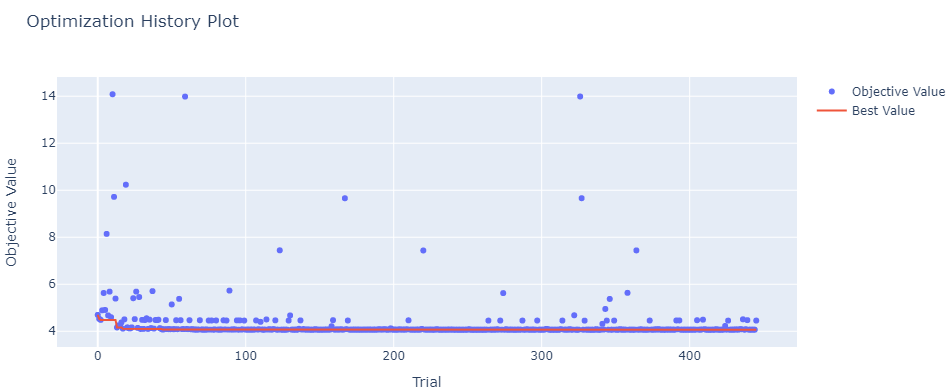
Tweedie regression

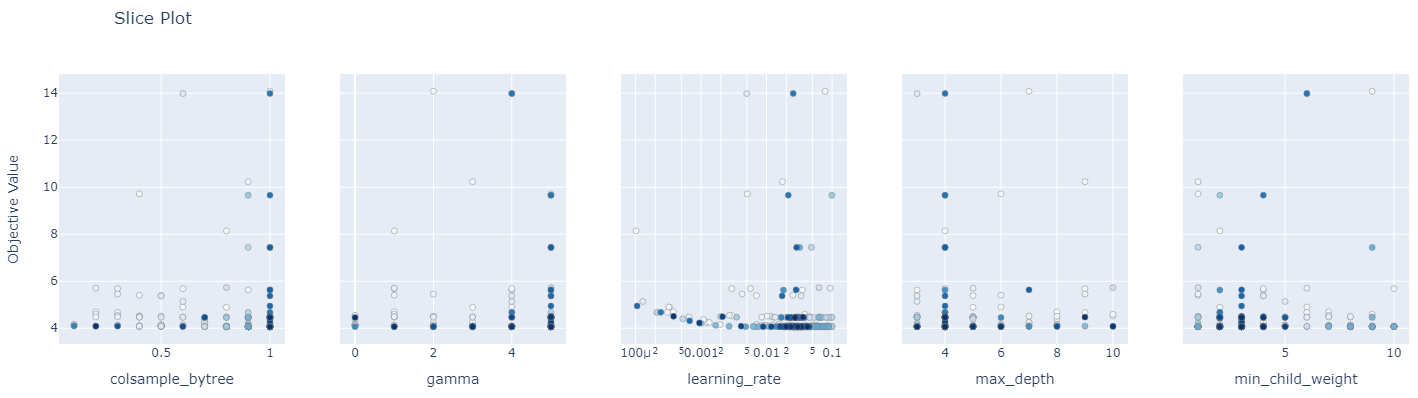
### Best hyperparameters:

Best is trial 413 with value: 4.065239666666667.

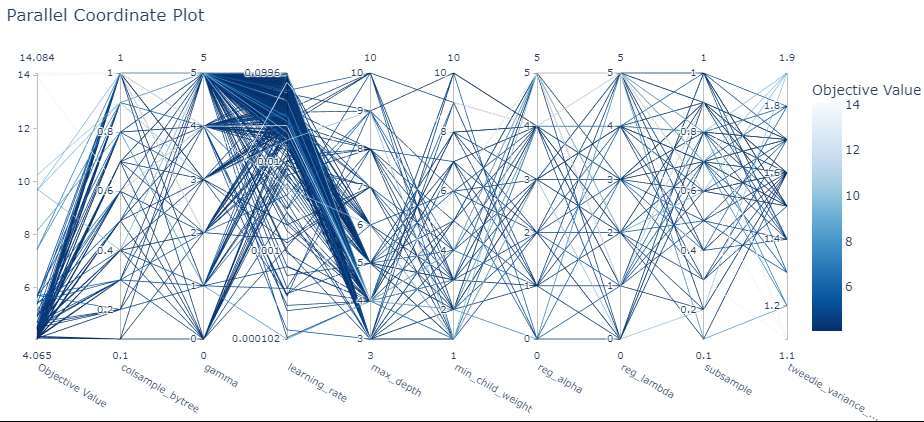
Study statistics: Number of finished trials: 446 Number of pruned trials: 0 Number of complete trials: 446 Best trial: val\_avg\_loss= Value: 4.065239666666667 Params: tweedie\_variance\_power: 1.6 max\_depth: 4 learning\_rate: 0.02632114605475799 subsample: 0.7000000000000001 colsample\_bytree: 1.0 gamma: 5 reg\_alpha: 5 reg\_lambda: 5 min\_child\_weight: 2 eta: 0.01

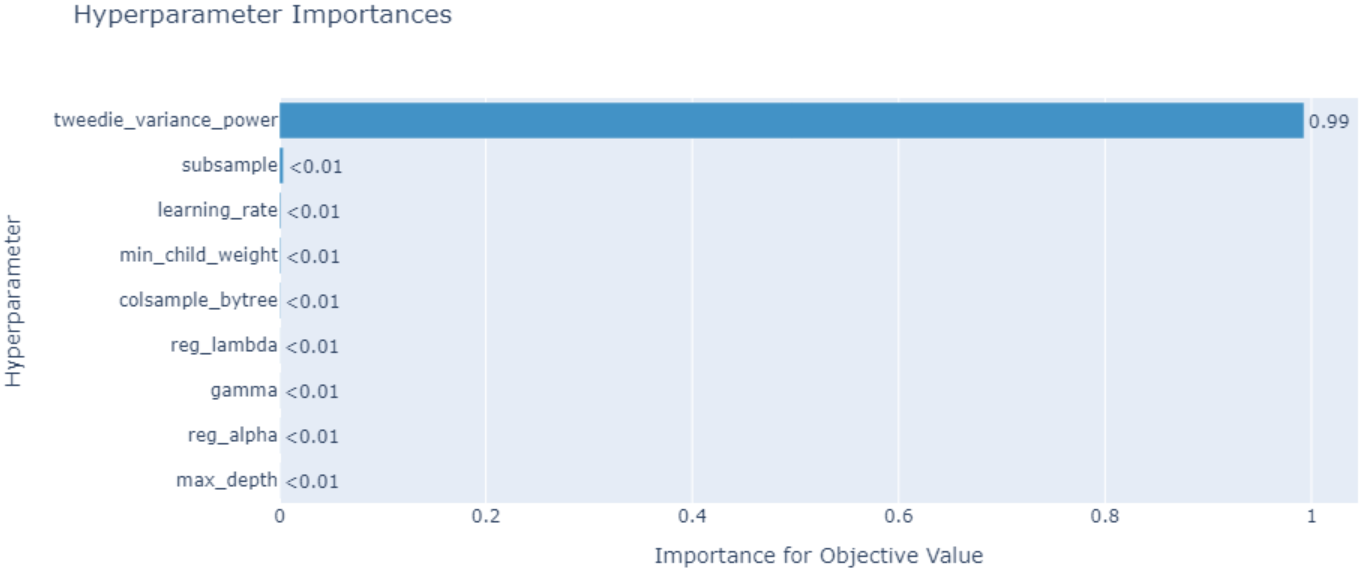
### Results:

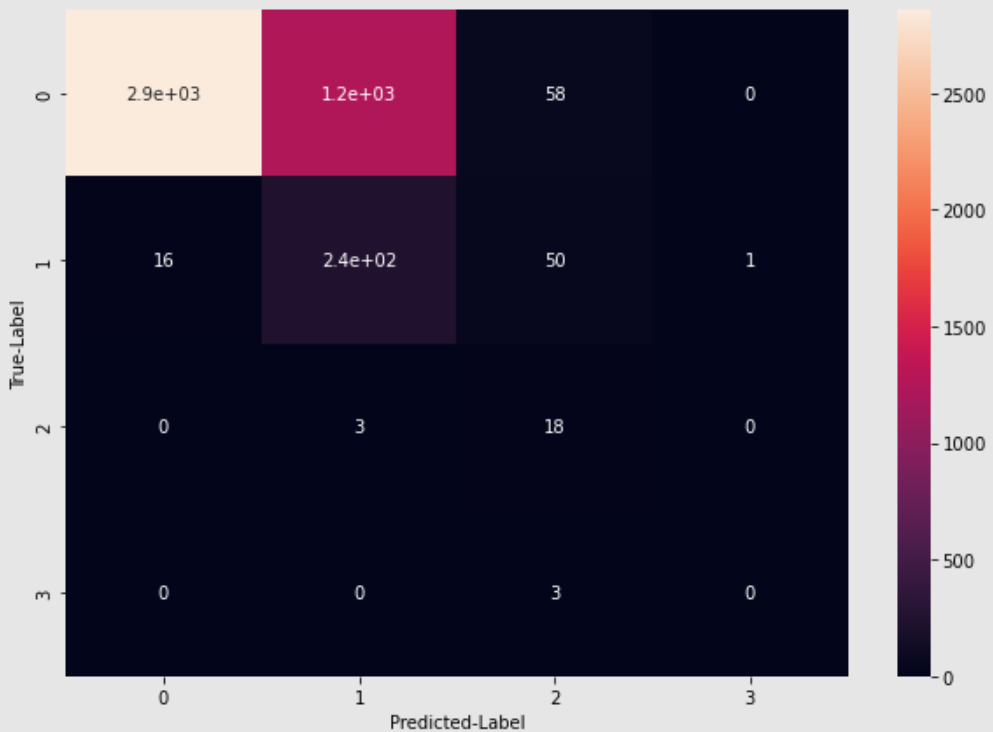




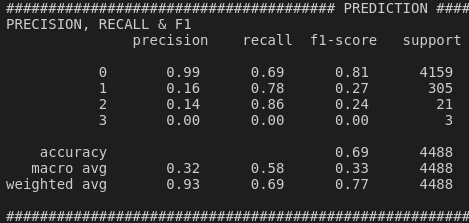








#### Precision,Recall & F1 score



### Analysis:

1) Predictions of cases 1,2 and 3 are OK. Improvement has to be made for class 0.

2) Need to include feature importance plot.

## Trial #2: Cross-entropy classification

### Dataset used:

### Features Importance:

### Loss:

* 1. Objective = multi:softmax

### Best hyperparameters:

Best trial:

Value: 0.7713229566566825

Params:

max\_depth: 3

eta: 0.055988067503231154

subsample: 0.30000000000000004

colsample\_bytree: 0.8

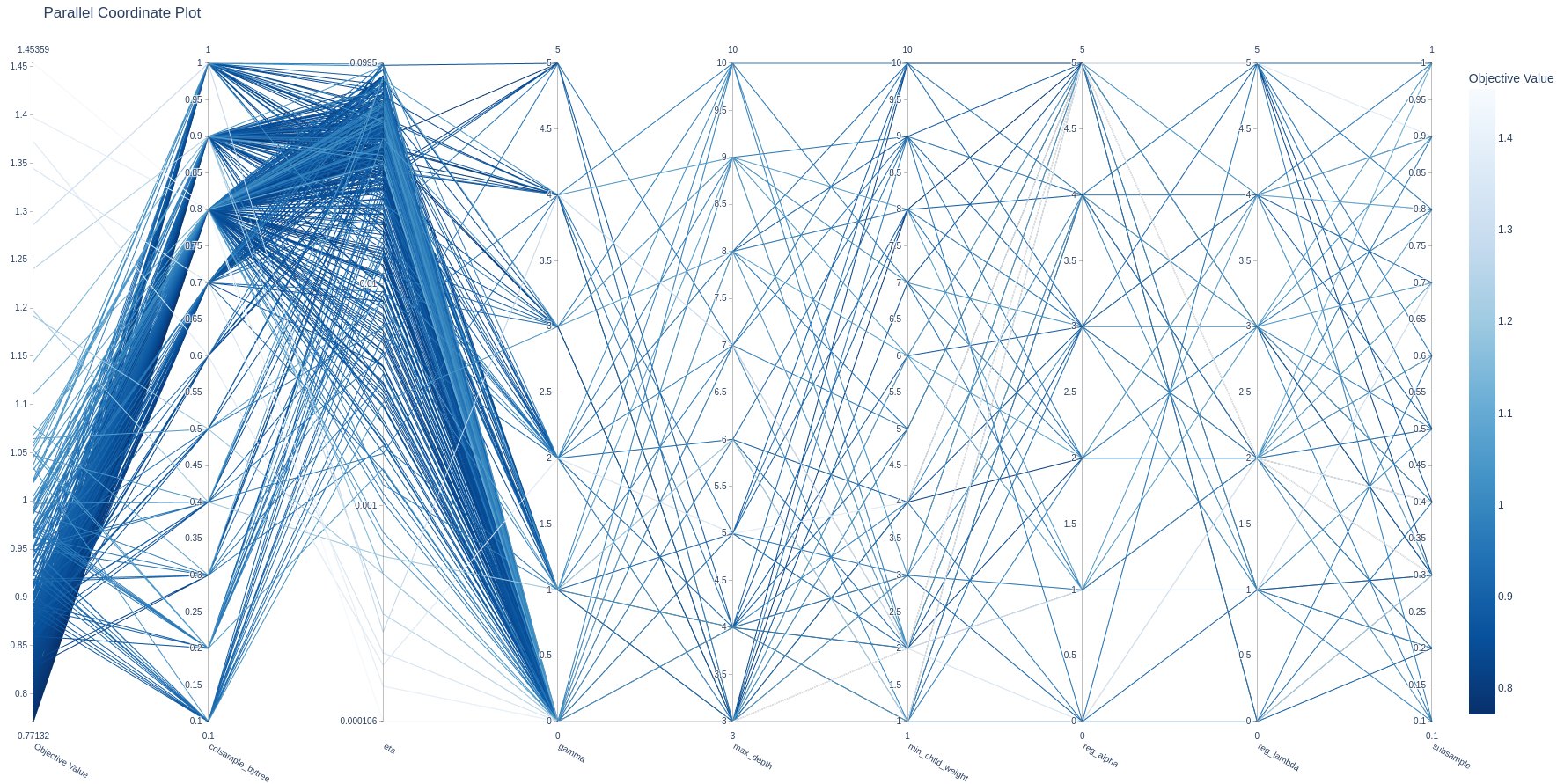
gamma: 0

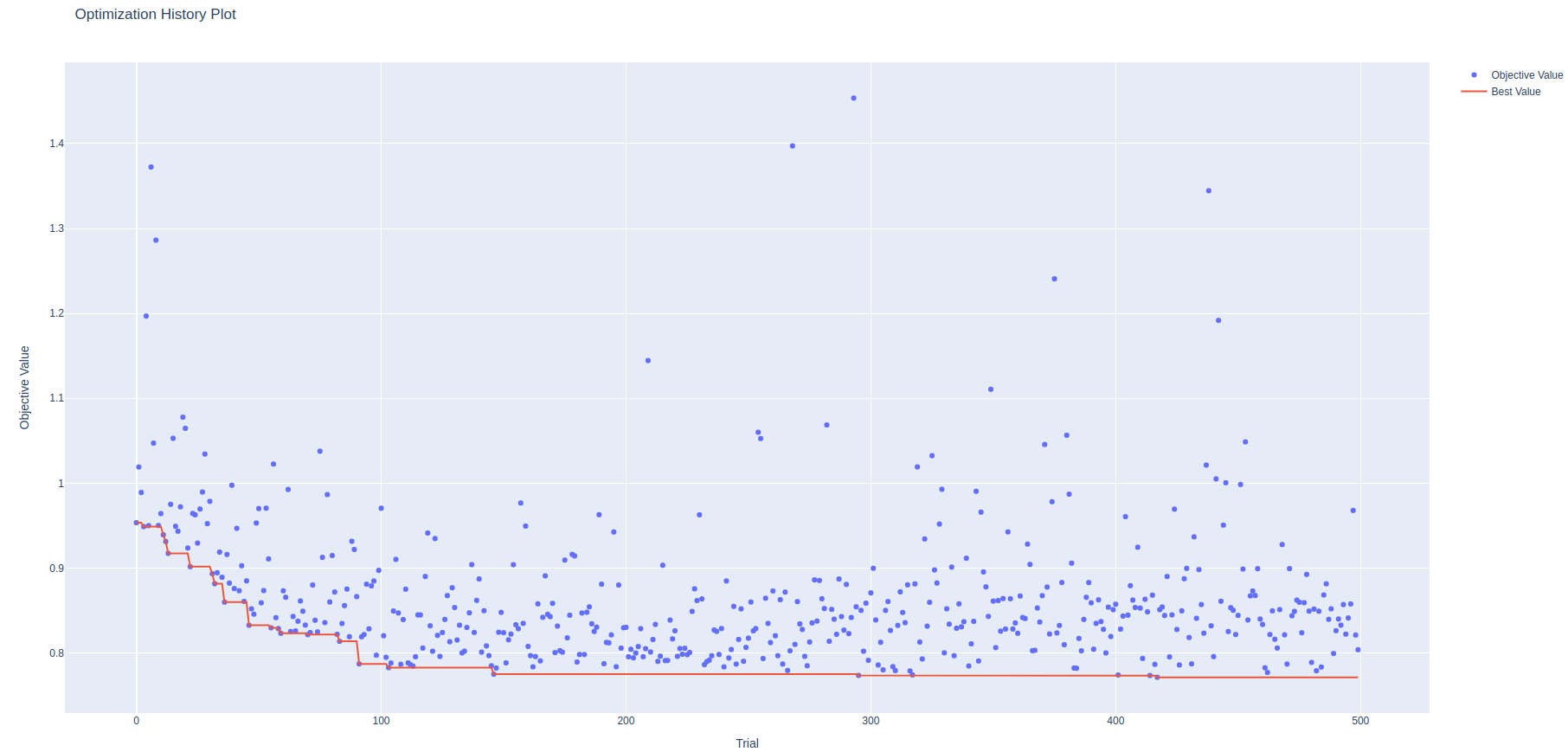
reg\_alpha: 5

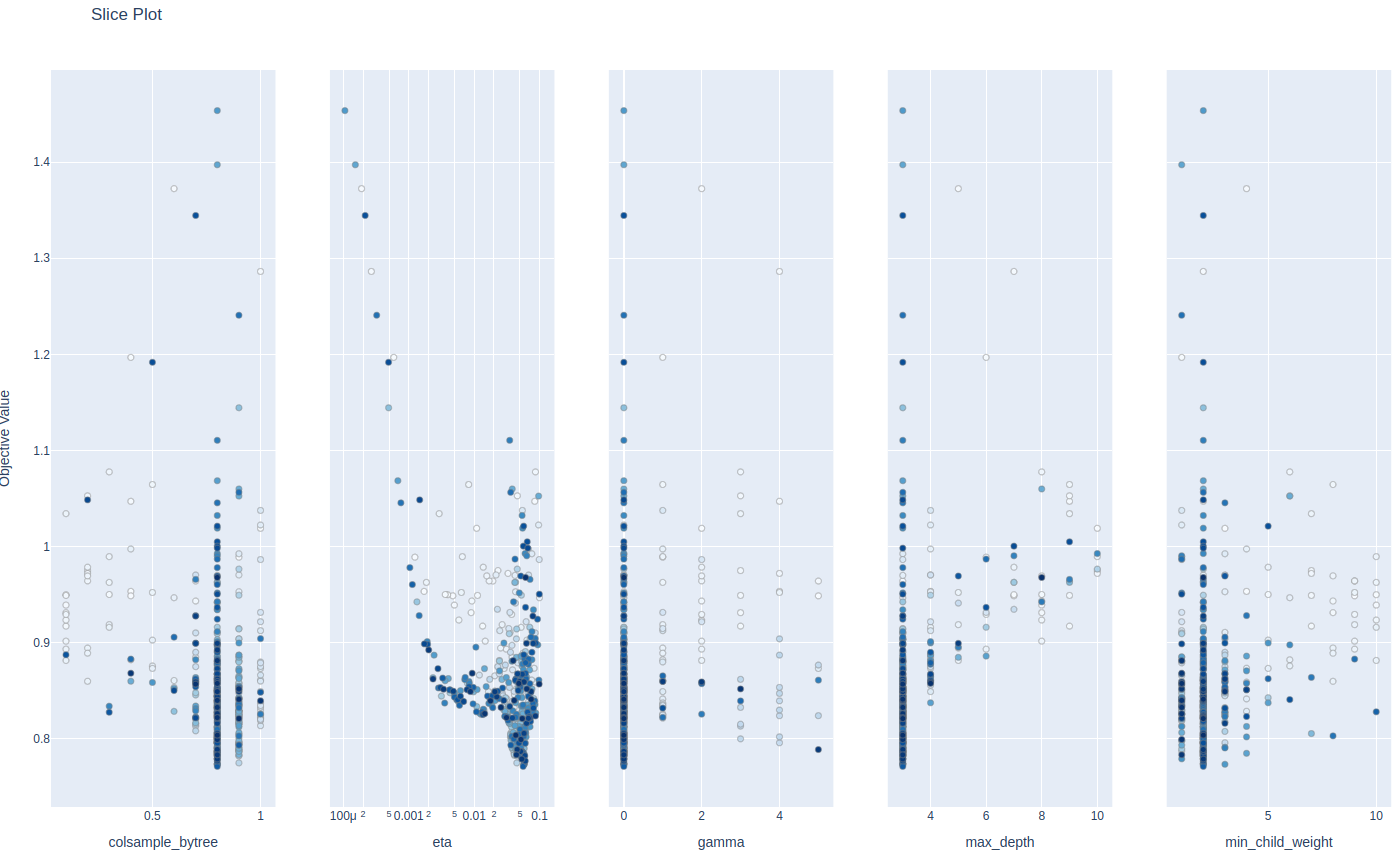
reg\_lambda: 2

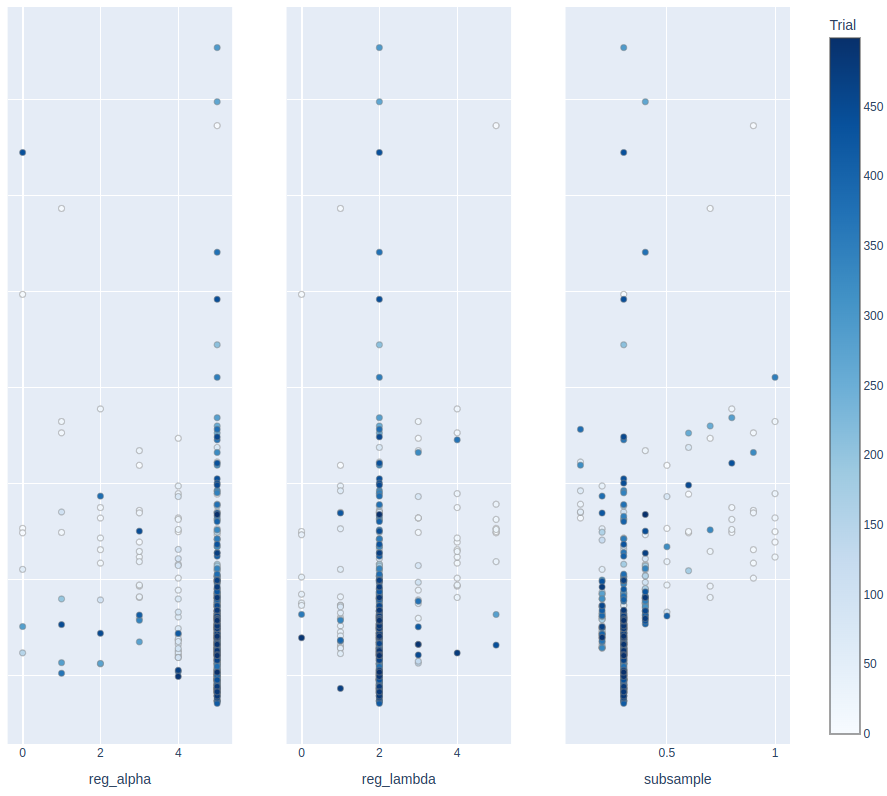
min\_child\_weight: 2

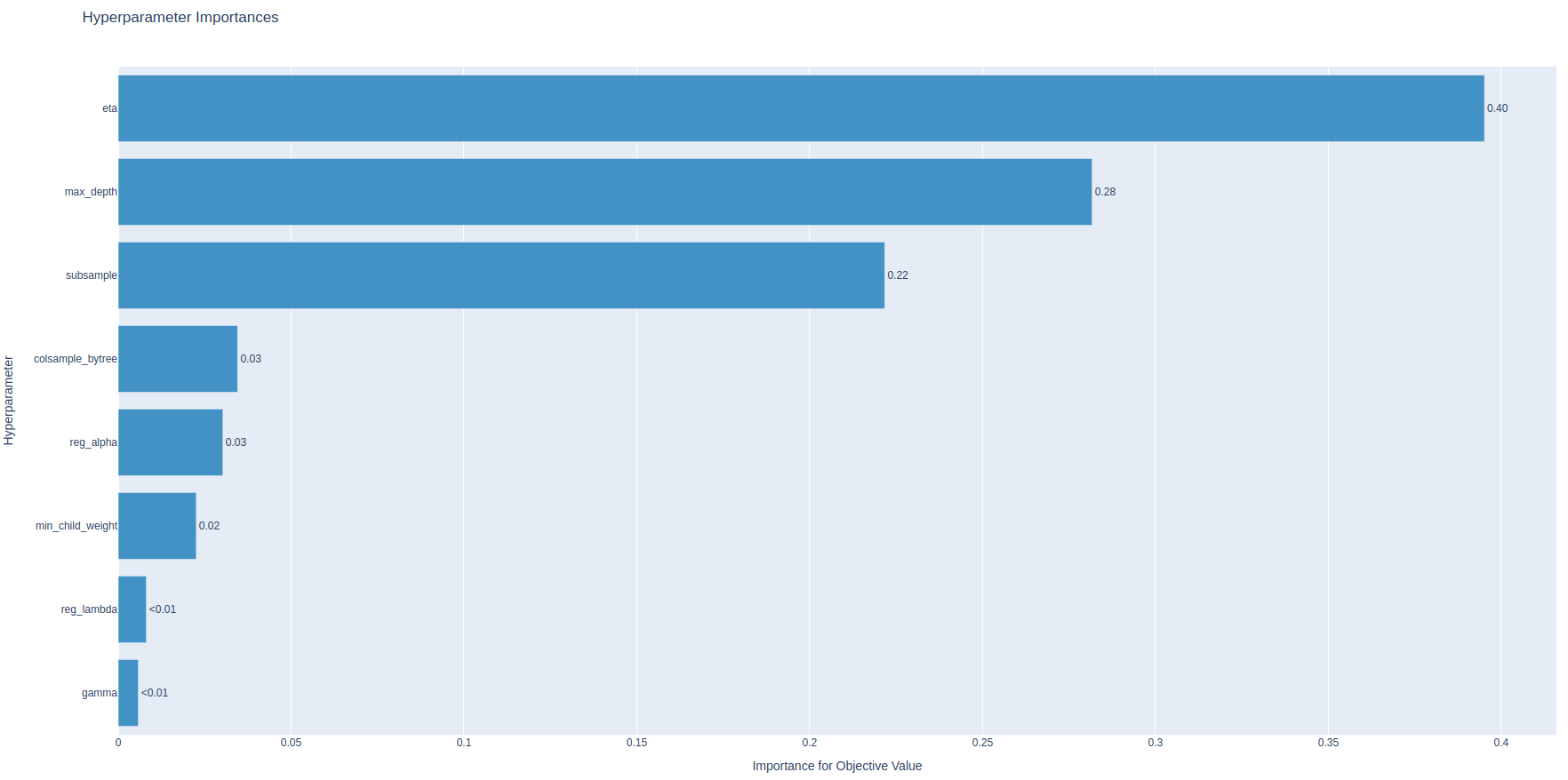
### Results:

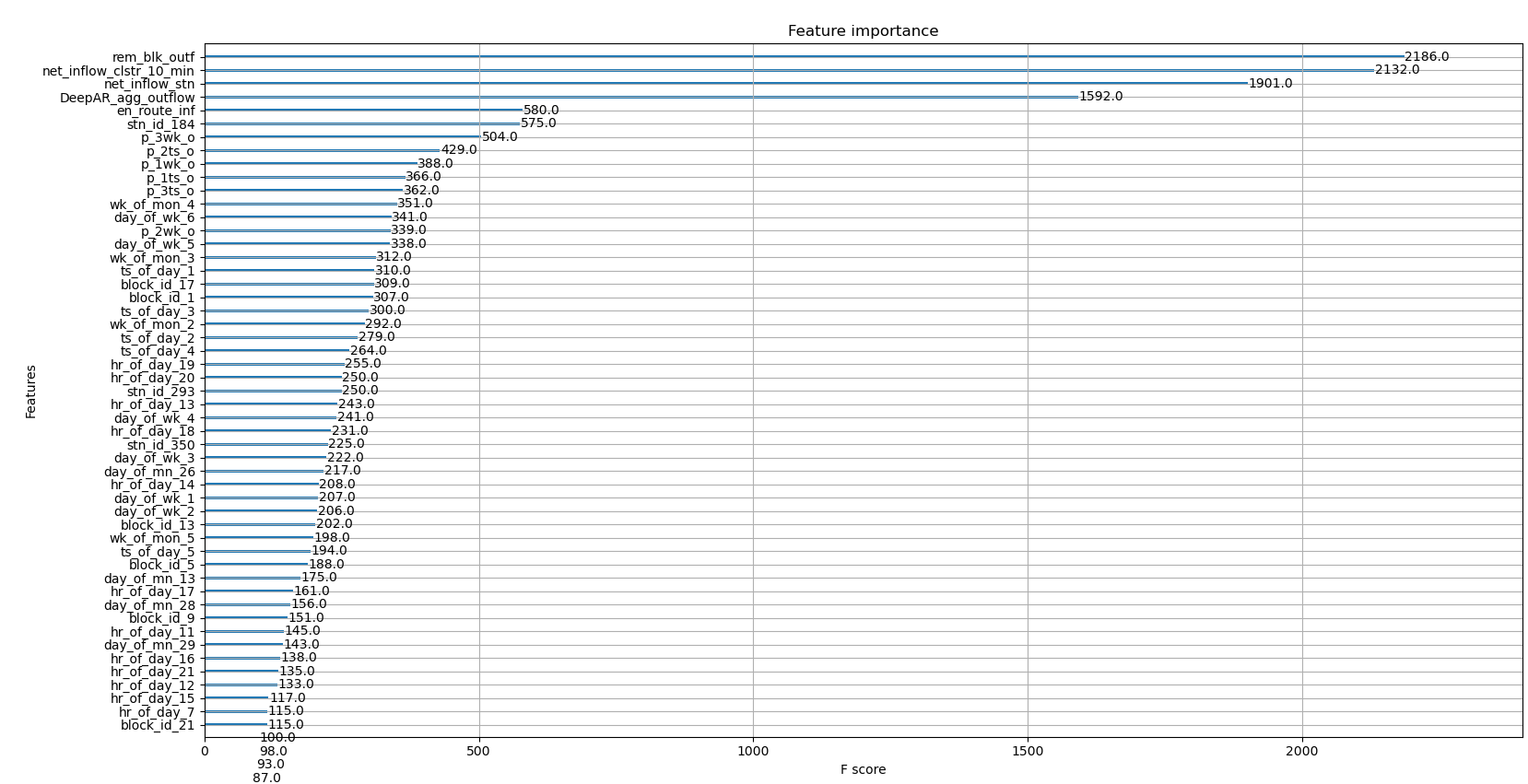


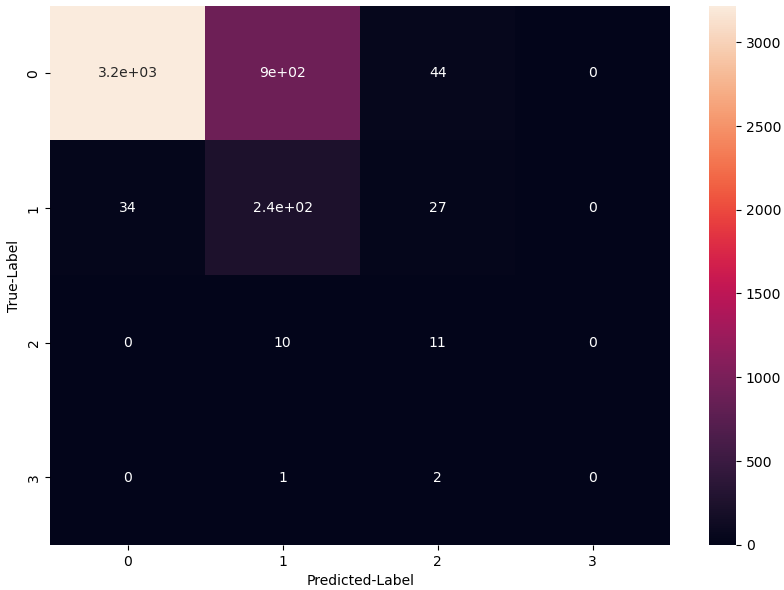




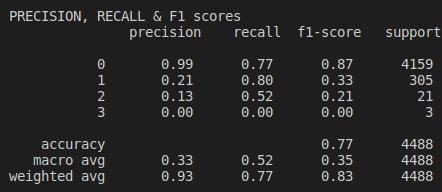








#### Precision,Recall & F1 score



### Analysis:

## Trial #3: More features and more region level dataset, Tweedie regression

## Trial #4: More features and more region level dataset, Cross-entropy classification.